

Домашнее задание 1 по базовой визуализации данных в R

Информационная часть

Дата выдачи: 20.09.2025

Дата дедлайна: **05.10.2025** 23:59 по Мск (GMT +3) – при всех расхождениях верны

эти дата и время

Максимально возможный балл: 25

При всех расхождениях верной считается информация из данного pdf-файла.

В задании вы продолжите работать с датасетом hogwarts. Данные, заполненные qmd-файлы с лекций-семинаров, а также пустой qmd для выполнения домашнего задания можно найти в папке на гугл-диске.

Полученный qmd-файл и результат knit в формате html загрузите на свой гитхаб. Папка, в которой лежат файлы с результатами домашнего задания на гитхабе должна называться basic_datavis_hw1_2025. Скрипт с домашним заданием будет называться BI_retarining_datavis_hw_1.qmd (доступен для скачивания на гугл-диске и в гугл-класссах). Все пути внутри проекта должны быть относительными. Папка data с данными должна лежать внутри папки, в которой находится ваш R-проект с домашним заданием (пути для чтения должны выглядеть таким образом: "data/kakoy_to_dataset.csv"). Чтобы сдать задание, приложите ссылку на гитхаб в гугл-классы. **Не забудьте добавить в коллабораторы на гитхабе Asklepiad (Богдан Сотников) и JuliGen (Юлия Нечаева), иначе мы не сумеем проверить вашу работу.**

```
basic_datavis_hw1_2025/  
├── BI_retarining_datavis_hw_1.qmd  
├── data  
│   ├── hogwarts_2025.csv  
│   └── podskazka.csv  
├── images  
└── Домашнее задание 1 по базовой визуализации данных в R.pdf
```

Пример структуры папки с домашним заданием.

Для этой (и следующей домашки) почти соблюдаются принципы классицизма: одно задание = не менее одного графика (и не менее одного чанка), если не указано иное. Это справедливо даже для ситуаций, когда следующее задание просит вас обратиться к предыдущему графику. В то же время, вы всегда можете расширить и дополнить свою интерпретацию или предложить дополнительные способы визуализировать то же самое другим, более подходящим, на ваш взгляд, способом, создав дополнительные графики для задания.

Некоторые пункты задания отсылают к аспектам, не упоминавшимся явно на лекции, и требуют обращения к справке и документации. В большинстве случаев, подобные моменты прописаны эксплицитно.

Если у вас возникают трудности технического, коммуникативного или

иного характера – можете писать в чат курса в телеграме или в личные сообщения @asklepiad (Богдан) @JuliN_Di (Юлия)

Задания

Во всех заданиях, пожалуйста, выполняйте несколько обязательных правил, даже если они не упомянуты в тексте задания явно.

- Делайте отображаемый текст на рисунке настолько крупным, насколько это позволяет сделать график и ваше чувство меры. Читаемые надписи – важный (но не очень часто встречающийся) элемент графиков.
- Не используйте в работе фон `ggplot2` по умолчанию.
- У каждого графика должен быть заголовок, отражающий его суть, корректно называющиеся оси.
- Старайтесь, чтобы графики отображались в достаточном крупном для читабельности размере не только в

срендеренном html, но и в процессе интерактивного исполнения внутри qmd.

Для выполнения некоторых заданий вам потребуется пакет `patchwork`, который мы уже загрузили ранее. Рекомендуем ознакомиться с [документацией](#) и использовать его в домашней работе.

Если при выполнении задания и интерпретации графика у вас возникает гипотеза, не стесняйтесь ее проверить – сделать еще один (или несколько) графиков для того же задания, провести дополнительные исследования и т.д.

Собственно задания

1. Постройте барплот (столбчатую диаграмму), отражающую распределение числа студентов по курсу обучения.
Раскрасьте столбики любым понравившимся вам цветом (можно использовать как словесные обозначения, так и гекскоды).
Добавьте цвет контура столбиков.
(1 б.)
2. Создайте новый барплот, отражающий распределение числа студентов по факультету. Добавьте на график вторую факторную переменную – происхождение (`bloodStatus`).
Оформите текст графика на русском языке.
Модифицируйте график при помощи аргумента `position` так, чтобы каждый столбец показывал распределение факультета по чистоте крови в долях.
Запишите текстом в qmd-документе, какой вывод можно сделать из графика?
(2 б.)
3. Отобразите распределение баллов, заработанных студентами на 3-й неделе обучения, по факультетам. Постарайтесь визуализировать одновременно и форму распределения, и некоторые описательные статистики. Отсортируйте факультеты в порядке убывания медианного балла за 3-ю неделю (мы не останавливались на этом в лекции, но упомянутая в ней функция по умолчанию сортирует именно по медиане, так что в этом случае дополнительных аргументов передавать не следует).

Если вы используете `geom_jitter`, постарайтесь, чтобы точки лежали кучнее на оси, не отражающей собственно распределение.
Проинтерпретируйте график.
(2 б.)

4. Постройте гистограмму для результата любого выбранного вами экзамена, кроме зельеварения.
Настройте оптимальное, на ваш взгляд, число столбцов гистограммы.
Выполните фасетирование по курсу. Постарайтесь, чтобы график был по возможности компактным.
(1 б.)
5. Постройте скаттерплот, визуализирующий связь между суммарным баллом студента за год и оценкой за экзамен по травологии.
Добавьте на график линию тренда. Удалите доверительную область и сделайте линию прямой.
Подумайте, как избежать того, чтобы записать одни и те же координаты x и y дважды.
Проинтерпретируйте график.
(1.5 б.)
6. Отобразите на одной иллюстрации скаттерплоты, аналогичные тому, что вы делали в предыдущем задании, для экзаменов по травологии, магловедению, нумерологии и зельеварению.
На иллюстрации для каждого из четырех графиков также должна присутствовать линия тренда с характеристиками, аналогичными тем, что были в предыдущем пункте.
Раскрасьте точки в разные цвета, в соответствии с факультетами.
Используйте стандартные цвета факультетов (как в лекционных `qmd`).
Проинтерпретируйте полученный результат.
(2 б.)
7. Видоизмените график, полученный на предыдущем шаге. Сгруппируйте и покрасьте линии тренда в соответствии с одной из категориальных переменных (с такой, которая подсвечивает одно из наблюдений на предыдущем этапе, относящееся ко всем 4-м экзаменам).
Постарайтесь избежать коллизий в легенде, при этом сохранив и цветовую палитру для факультетов. Проинтерпретируйте результат.

(2 б.)

8. Сравните распределение баллов за экзамен по прорицанию и за экзамен по защите от темных искусств. Сделайте это тремя разными способами. Под разными способами понимаются идеологически разные геомы или способы группировки. Не считаются разными способами изменения константных визуальных параметров (цвет заливки, размер фигур, другие чисто стилевые отличия) на сходных в остальном графиках.

Объедините графики, таким образом, чтобы результирующий график имел два столбца и 2 строки. Два графика должны находиться в верхней строке и занимать равную площадь. Третий график должен занимать нижнюю строку целиком. Проинтерпретируйте график.

(2 б.)

9. Визуализируйте средний балл по зельеварению студентов с различным происхождением.

Вы вольны добавить дополнительные детали и информацию на график, например, какую-то из мер рассеяния и текстовые элементы.

Измените порядок ваших фигур слева направо следующим образом: маглорожденные, полукровки, чистокровные.

Проинтерпретируйте результат. Как вы думаете, почему он именно такой? Если у вас есть гипотеза, проиллюстрируйте ее еще одним графиком (или графиками).

Если у вас возникают сложности, обратитесь к шпаргалке по пакету `forcats` от `posit`.

(2 б. за первый график с правильной интерпретацией и 2 б. за подтверждение в виде последующих графиков и текстовой аргументации. Итого 4 б.)

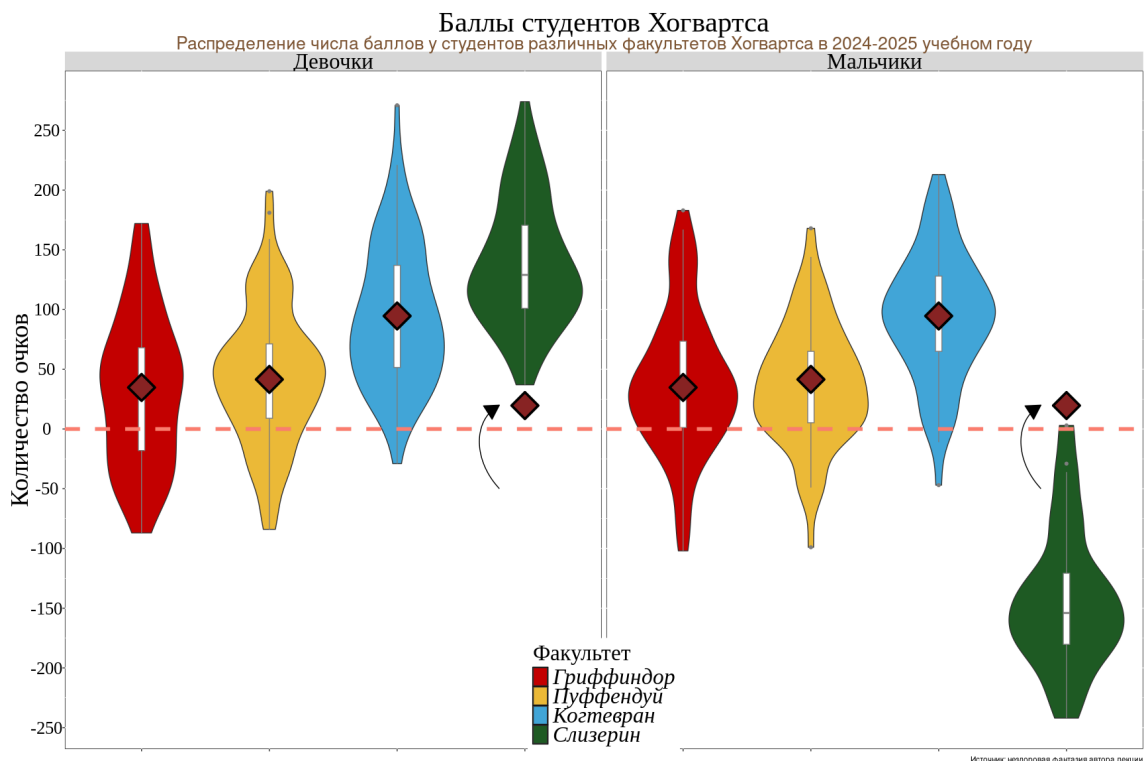
10. Изобразите на графике диаграмму рассеяния (скаттерплот), где на осях располагались бы результаты экзаменов по заклинаниям и по древним рунам.

На графике присутствует некоторая проблема. Если вы затрудняетесь ее идентифицировать, то можете взять подсказку -- для этого постройте скаттерплот для датасета ``podskazka.csv``, лежащего в папке ``data`` архива из домашнего задания -- по оси `x` отобразите переменную ``y``, а по оси `y` -- переменную ``x`` и с настройкой лимитов по оси `y` от 200 до 500.

Предложите и подкрепите визуализациями не менее 3-х способов решения означенной проблемы.
(2 б.)

11. Используйте датасет с набором баллов по неделям для первых 30-и студентов, который мы получили на лекции. Постройте лайнплоты для всех студентов. Выделите при помощи пакета `gghighlight` студентов с факультета Когтевран, на лейблах отразите их происхождение. Линии, не относящиеся к студентам-когтевранцам покрасьте в небесно-голубой цвет, толщину их настройте на уровне не более трети от стандартной.
(1 б.)

12. Воспроизведите график максимально близко к оригиналу и проинтерпретируйте его.
Постарайтесь максимально точно отразить каждый элемент.
(4.5 б.)



Полезные ссылки

1. [Шпаргалка по ggplot2](#)
2. [Шпаргалка по dplyr](#)
3. [Шпаргалка по forcats](#)
4. [Документация ggplot2](#)
5. [Документация patchwork](#)
6. [Документация gghighlight](#)